Predicting Day-Ahead MISO's Locational Marginal Prices Using Data Mining Techniques and Publicly Available Data

LaRico Andres
Michael Acquah
CECS
ECE 537

DEARBORN

Contributions

LaRico Andres

- Data Collection
- Preprocessing
- Background Research
- Presenting and Reporting

Michael Acquah

- Feature Engineering
- Model Development
- Project Planning and Coordination

Both

- Documentation
- Visualization
- Communication
- Literature Review

What is Locational Marginal Pricing (LMP)

Locational Marginal Pricing (LMP) is a pricing mechanism used in electricity markets to reflect the cost of delivering power to specific locations, or nodes, within a transmission network. It accounts for the cost of electric power generation, the cost of delivering that power, and the physical limitations of the transmission system. LMP is crucial in managed wholesale markets, providing real-time pricing signals that help balance supply and demand while considering factors like congestion and load patterns. The Federal Energy Regulatory Commission (FERC) supports LMP as it promotes efficiency in wholesale electricity markets

Problem Statement

This project intends to apply data mining techniques such as data preprocessing:

- Transformations
- Correlations
- Normalizations

Time variant warehousing:

Data	How
Timestamp	Every row is tied to an hour
LMP values (log_LMP)	Derived from real hourly data
Weather Forecasts	Hourly and date-specific forecasts
Load Forecasts	Provided by hour/date
Lag Features (LMP_lag_1, LMP_lag_24)	Capture temporal dependency over time

Publicly Available Data:

- US Energy Information Administration (EIA) **LMP Data**
- Open-Meteo.com **Weather API**
- Midcontinent Independent System Operator (MISO) **Load Data**

Machine Learning Models Used:

Model	Type	Purpose
Linear Regression	Supervised	Baseline Model LMP
Random Forest	Supervised	Primary Forecasting and Reconstruction
XGBoost	Supervised	Benchmarking
Kmeans	Unsupervised	Clustering time/weather/LMP patterns
PCA	Unsupervised	Dimensionality reduction

Initial LMP Data
- Timestamp
- Hour Number
- Michigan Hub LMP
- Michigan Hub (Congestion)
- Michigan Hub (Loss)

Understanding Data
- Check Variable Type
- Get Descriptive Statistics
- **IQR Outlier Detection**
- Get Initial Visualization

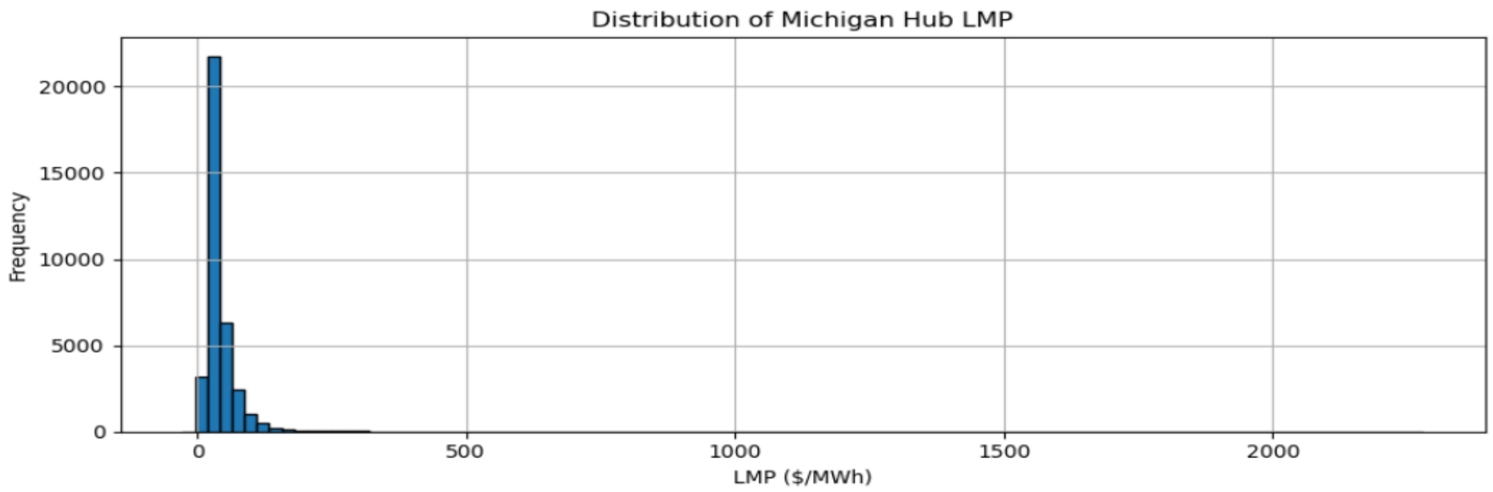
Data Preparation

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Timestamp                  35944 non-null  datetime64[ns]
 1   Hour Number                35944 non-null  int64
 2   Michigan Hub LMP           35944 non-null  float64
 3   Michigan Hub (Congestion)  35944 non-null  float64
 4   Michigan Hub (Loss)        35944 non-null  float64
dtypes: datetime64[ns](1), float64(3), int64(1)
memory usage: 1.4 MB
None
```

```
📊 Descriptive Statistics:
        Hour Number  Michigan Hub LMP  Michigan Hub (Congestion)
count   35944.00000      35944.000000               35944.000000
mean       12.49744         41.277227                   0.895211
std         6.92239         41.026586                  11.364288
min         1.00000        -27.470000                -405.830000
25%         6.00000         23.447500                   0.000000
50%        12.00000         30.660000                   0.000000
75%        18.00000         46.970000                   0.950000
max        24.00000       2280.330000                 373.380000

       Michigan Hub (Loss)
count         35944.000000
mean              1.046162
std               1.931971
min             -48.130000
25%               0.320000
50%               0.840000
75%               1.540000
max              84.580000
```

```
❗ Missing Values:
Timestamp                    0
Hour Number                  0
Michigan Hub LMP             0
Michigan Hub (Congestion)    0
Michigan Hub (Loss)          0
dtype: int64
```

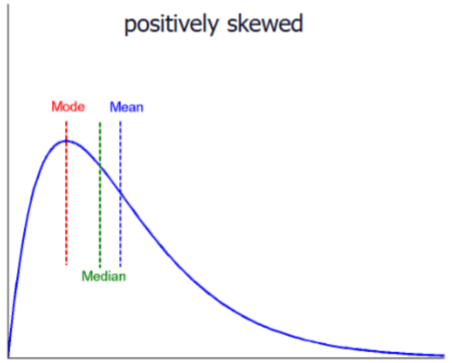


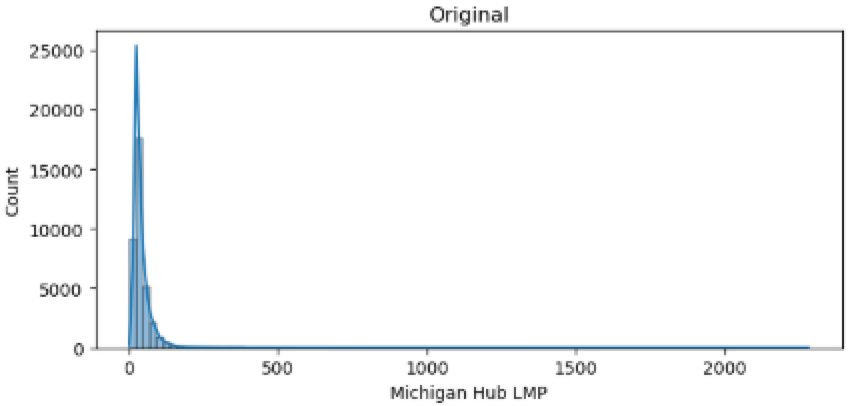
Graphic Display:
Histogram

IQR Analysis Data Mining Technique #1

We need to better understand our initial dataset: We have observed our data is positively-skewed.

```
Mean: 41.28
Median: 30.66
Mode: 22.34
```






```python
# Step 1: Calculate Q1 and Q3
Q1 = df['Michigan Hub LMP'].quantile(0.25)
Q3 = df['Michigan Hub LMP'].quantile(0.75)
Median = df['Michigan Hub LMP'].quantile(0.50)
# Step 2: Compute IQR
IQR = Q3 - Q1

# Step 3: Define bounds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Step 4: Flag outliers
df['is_outlier'] = (df['Michigan Hub LMP'] < lower_bound) | (df['Michigan Hub LMP'] > upper_bound)
```

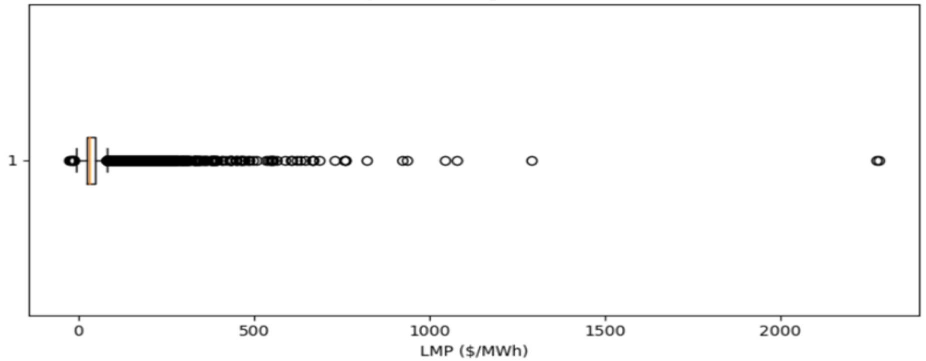
Outliers are values outside of the interval [lower_bound, upper_bound]

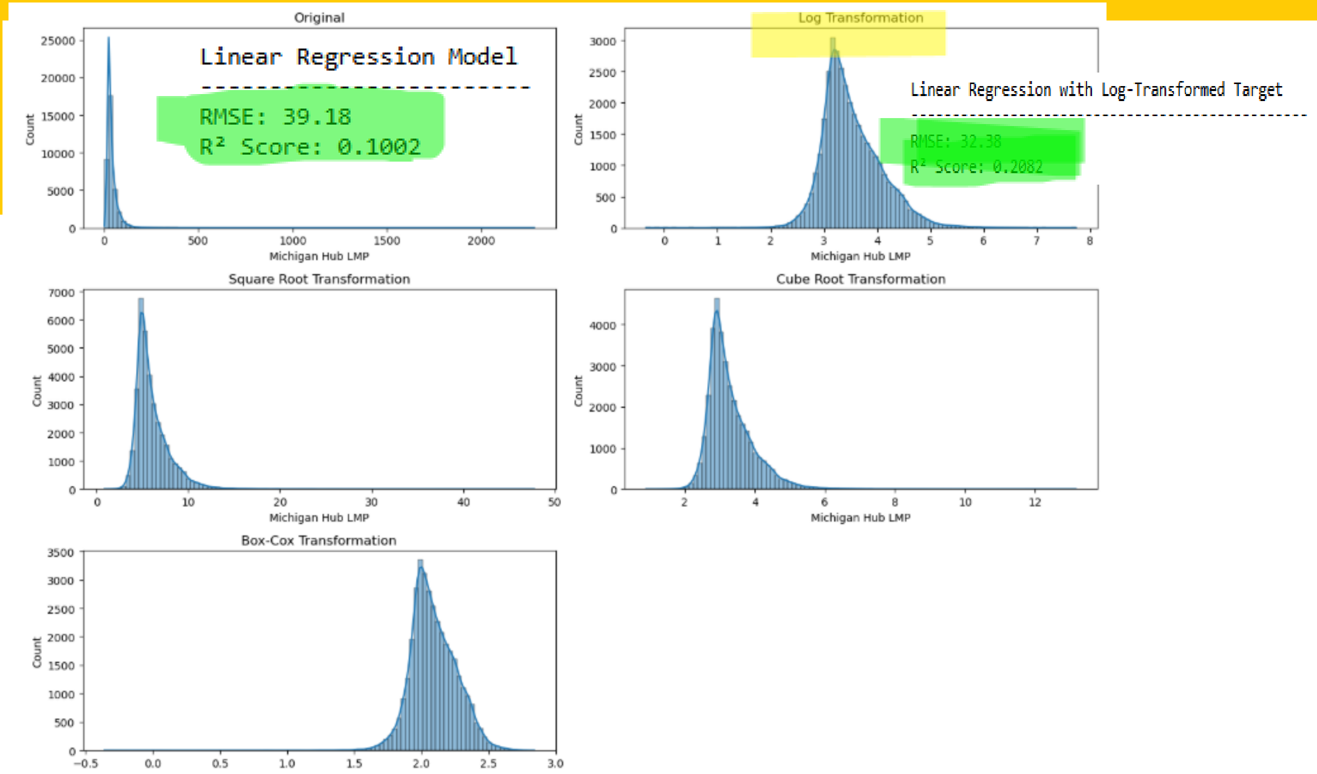
Five-number summary

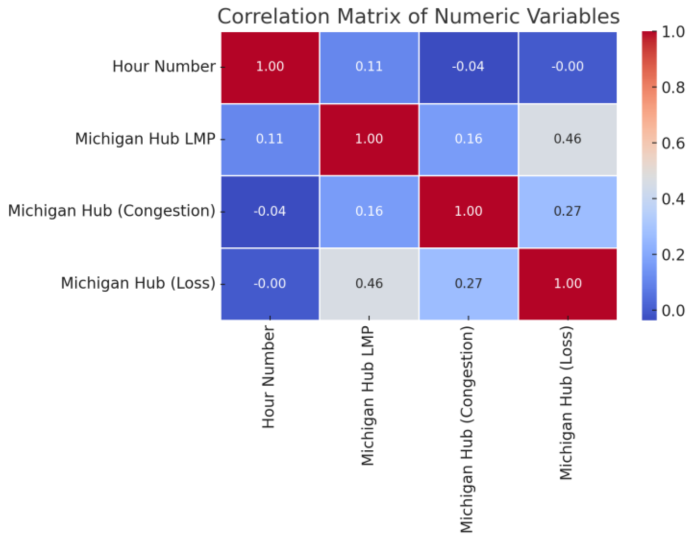
```
Q1: 23.447499999999998
Median: 30.66
Q3: 46.97
IQR: 23.5225
Total records: 35944
Outlier count: 2657
Outlier percentage: 7.39%
Lower bound: -11.84
Upper bound: 82.25
```

	Hour Number	Michigan Hub LMP
count	35944.00000	35944.000000
mean	12.49744	41.277227
std	6.92239	41.026586
min	1.00000	-27.470000
25%	6.00000	23.447500
50%	12.00000	30.660000
75%	18.00000	46.970000
max	24.00000	2280.330000

Boxplot of Michigan Hub LMP







Michigan Hub LMP has:

•**Strong positive correlation** with **Congestion** and **Loss** components.

•Weak correlation with **Hour Number** (as expected — time of day alone doesn't fully explain price changes).

- **Data Preparation**
- Handle Missing Values, duplicates, and outliers.
- Perform Feature Engineering
- Data Integration and Formatting
- Splitting Data

Plots

Original-Strongly positively-skewed, heavy tail

Log Transformation – Greatly reduces skewness; commonly used for price data

Square Root – Milder effect, put still pulls in outliers

Cube Root – Useful for handling large range values

Box-Cox – Automatically chooses the best* exponent.

Final LMP Data

- Timestamp
- Hour Number
- Michigan Hub LMP
- Michigan Hub (Congestion)
- Michigan Hub (Loss)
- DayOfWeek
- Month
- IsWeekend
- Temperature_2m
- Relative_humidity_2m
- dew_point_2m
- Preciitation
- Rain
- Snowfall
- Snow__depth
- Weather_code
- Wind_speed_10m
- Wind_direction_10m
- Wind_gusts_10m
- Actual_load

Understanding Data

- Check Variable Type
- Get Descriptive Statistics
- **IQR Outlier Detection**
- Get Initial Visualization

Data Preparation

DEARBORN

#	Column	Non-Null Count	Dtype
0	Timestamp	33852 non-null	object
1	Hour Number	33852 non-null	int64
2	Michigan Hub LMP	33852 non-null	float64
3	Michigan Hub (Congestion)	33852 non-null	float64
4	Michigan Hub (Loss)	33852 non-null	float64
5	Hour	33852 non-null	int64
6	DayOfWeek	33852 non-null	int64
7	Month	33852 non-null	int64
8	IsWeekend	33852 non-null	int64
9	temperature_2m	33852 non-null	float64
10	relative_humidity_2m	33852 non-null	float64
11	dew_point_2m	33852 non-null	float64
12	precipitation	33852 non-null	float64
13	rain	33852 non-null	float64
14	snowfall	33852 non-null	float64
15	snow_depth	33852 non-null	float64
16	weather_code	33852 non-null	int64
17	wind_speed_10m	33852 non-null	float64
18	wind_direction_10m	33852 non-null	float64
19	wind_gusts_10m	33852 non-null	float64
20	Actual_load	33852 non-null	float64

📊 Summary Statistics:

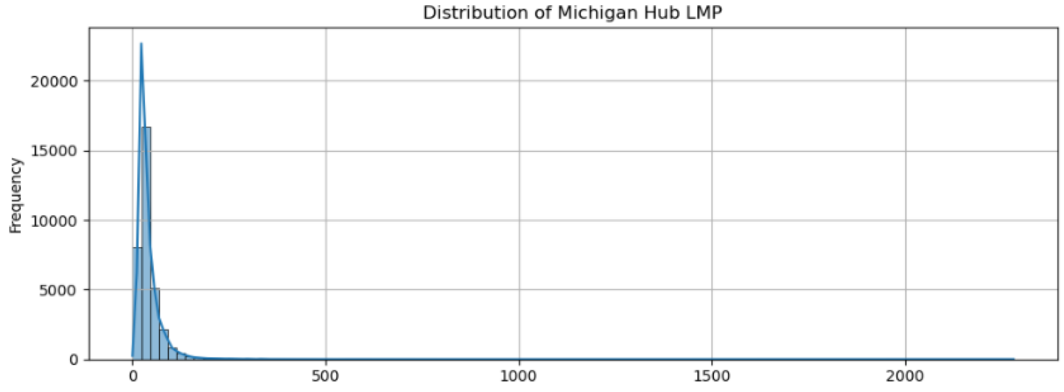
	Hour Number	Michigan Hub LMP	Michigan Hub (Congestion)
count	33852.000000	33852.000000	33852.000000
mean	12.501152	42.052714	0.946226
std	6.922327	41.511380	11.653210
min	1.000000	0.710000	-405.830000
25%	7.000000	23.777500	0.000000
50%	13.000000	31.425000	0.000000
75%	19.000000	48.112500	1.040000
max	24.000000	2280.330000	373.380000

	Michigan Hub (Loss)	Hour	DayOfWeek	Month
count	33852.000000	33852.000000	33852.000000	33852.000000
mean	1.077374	11.501152	2.996662	6.380037
std	1.967204	6.922327	1.997418	3.502063
min	-48.130000	0.000000	0.000000	1.000000
25%	0.340000	6.000000	1.000000	3.000000
50%	0.870000	12.000000	3.000000	6.000000
75%	1.580000	18.000000	5.000000	10.000000
max	84.580000	23.000000	6.000000	12.000000

	IsWeekend	temperature_2m	relative_humidity_2m	dew_point_2m
count	33852.000000	33852.000000	33852.000000	33852.000000
mean	0.284828	53.774666	68.763668	42.767090
std	0.451339	19.091454	17.792520	18.486589
min	0.000000	-10.312599	15.064703	-21.202599
25%	0.000000	39.187400	55.142455	29.647400
50%	0.000000	55.027397	69.802067	44.047400
75%	1.000000	69.247400	83.686082	58.357400
max	1.000000	95.347400	100.000000	81.037400

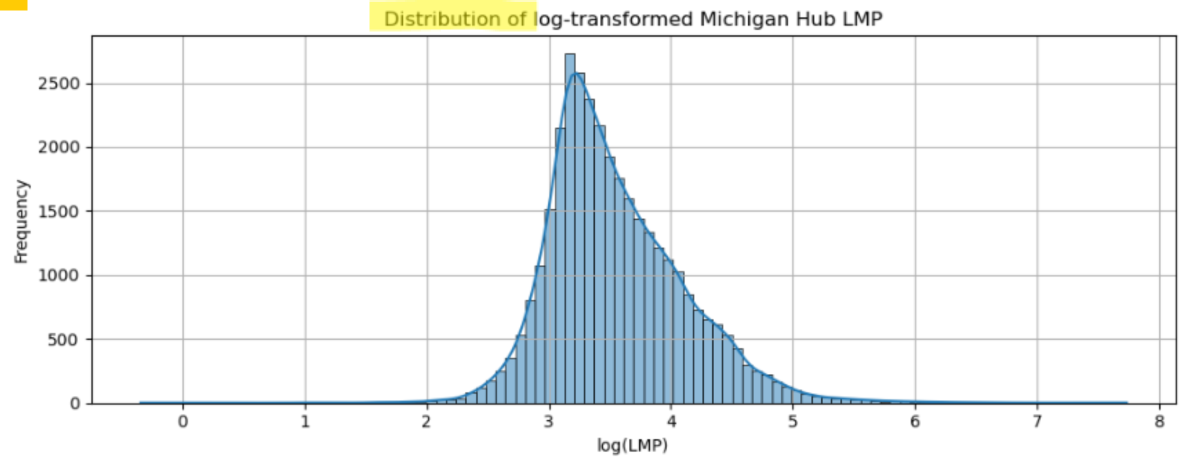
	precipitation	rain	snowfall	snow_depth	weather_code
count	33852.000000	33852.000000	33852.000000	33852.000000	33852.000000
mean	0.005020	0.004771	0.001745	0.023553	9.600142
std	0.025823	0.025635	0.020645	0.092691	19.516105
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	3.000000
75%	0.000000	0.000000	0.000000	0.000000	3.000000
max	0.818898	0.818898	0.799213	0.951444	75.000000

	wind_speed_10m	wind_direction_10m	wind_gusts_10m	Actual_load
count	33852.000000	33852.000000	33852.000000	33852.000000
mean	7.594532	193.433748	15.395790	18088.428771
std	3.700705	93.702219	7.305166	2998.426809
min	0.000000	0.535451	0.894800	12103.280000
25%	4.829019	128.659835	9.619101	15896.312500
50%	6.938307	200.462360	14.316800	17845.120000
75%	9.812248	268.830900	19.909300	19561.687500
max	29.747265	360.000000	56.819798	33064.270000

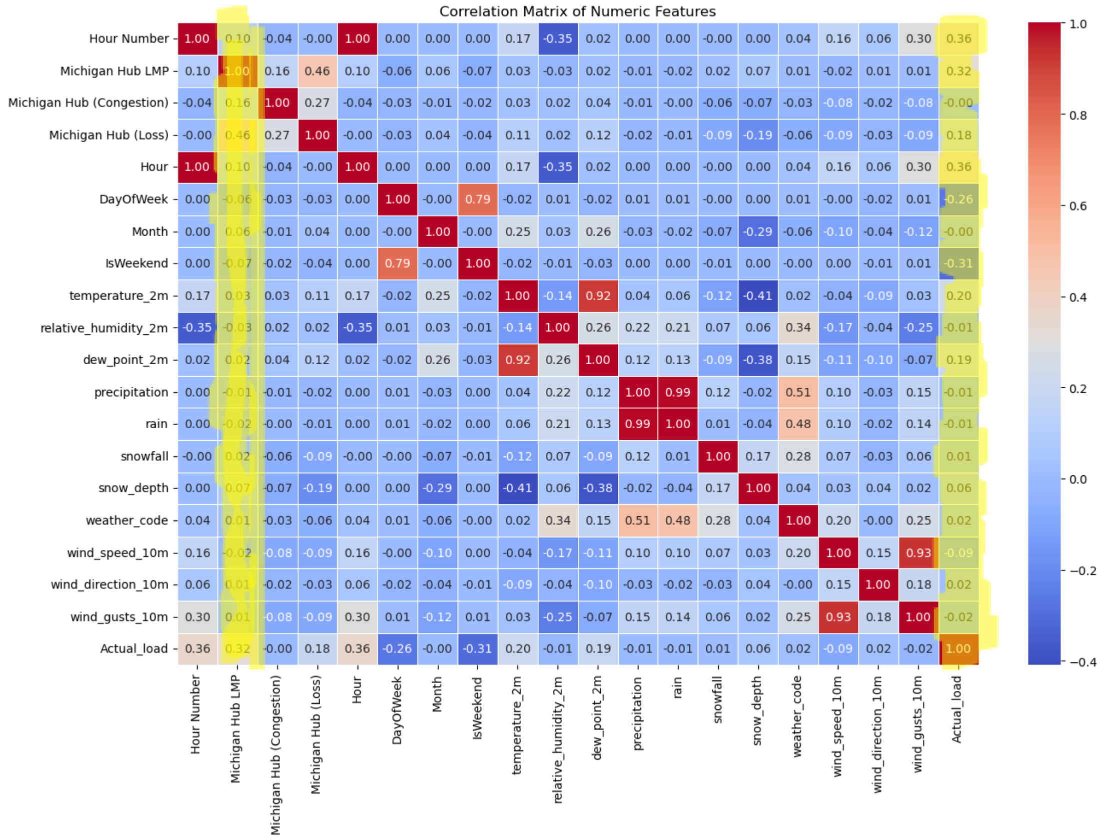


Graphic Display: Histogram

Log Transformation



Pearson Correlation Matrix



- **Data Preparation**
- Handle Missing Values, duplicates, and outliers.
- Perform Feature Engineering
- Data Integration and Formatting
- Splitting Data

Lessons From Correlation Matrix

Top Correlated Features with MLP:

- *Michigan Hub (Congestion) +0.46, Strongest positive correlation
- *Michigan Hub (Losses) +0.27, Losses in the system can affect LMP
- *Actual Load +0.32, Load affects supply-demand balance, higher demand higher LMP
- *Hour +0.10, LMP varies throughout the day correlation *Temperature_2m +0.03
- *Dewpoint_2m +0.07, Weak correlation but may still contribute
- *IsWeekend +0.07 Some weekend effect
- Hour and Hour Number are Multi-collinear, will delete one **REDUNDANT**
- Rain, snowfall, precipitation are correlated as well, will keep only one **REDUNDANT**

Next steps:

Select key features for model(*) FEATURE SELECTION
Build baseline model (simple regression, Linear or RandomForest)
Test non-linear models XGBoost Light GBM

Data Understanding

Attributes

	Timestamp	Hour Number	Michigan Hub LMP	Michigan Hub (Congestion)	Michigan Hub (Loss)	Hour	DayOfWeek	Month	IsWeekend	temperature_2m	...	rain	snowfall	snow_depth	weather_code	wind_speed_10m	wind_direction_10m	wind_gusts_10m	Actual_load	log_LMP	is_outlier
0	2021-02-10 00:00:00	1	24.43	0.00	-0.89	0	2	2	0	18.487400	...	0.0	0.0	0.360892	3	4.529580	32.905247	9.395399	18226.97	3.195812	False
1	2021-02-10 01:00:00	2	24.39	0.00	-1.01	1	2	2	0	18.397400	...	0.0	0.0	0.360892	3	4.787389	37.405437	9.395399	17785.44	3.194173	False
2	2021-02-10 02:00:00	3	24.38	0.00	-0.95	2	2	2	0	16.597400	...	0.0	0.0	0.360892	3	4.273782	47.121110	9.395399	17582.40	3.193763	False
3	2021-02-10 03:00:00	4	26.32	0.00	-0.99	3	2	2	0	14.977398	...	0.0	0.0	0.328084	3	4.412054	59.534540	8.276900	17527.62	3.270329	False
4	2021-02-10 04:00:00	5	30.73	0.00	-0.96	4	2	2	0	13.537399	...	0.0	0.0	0.328084	3	4.654895	54.782326	8.053200	17753.76	3.425239	False

Data Objects

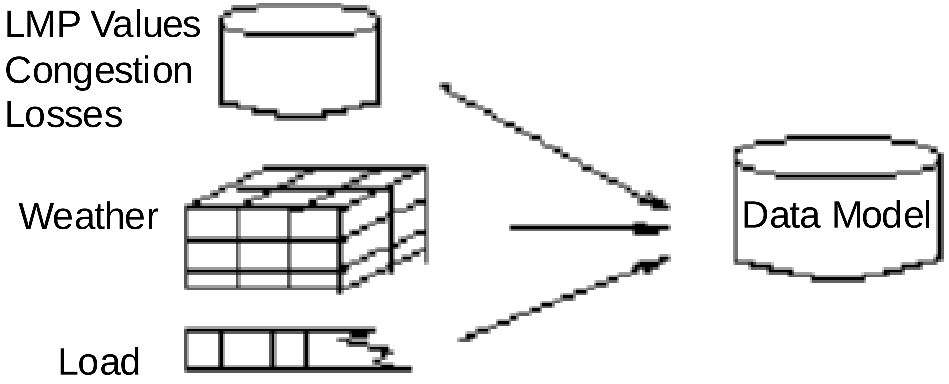
IsWeekend is Binary

Uniqueness Rule, Consecutive Rule and Null Rule

🔍 Data Quality Summary:

	Unique Values	Null Count	Consecutive Changes
Actual_load	33252	0	33852
DayOfWeek	7	0	1410
Hour Number	24	0	33852
IsWeekend	2	0	407
Michigan Hub (Congestion)	3608	0	21518
Michigan Hub (Loss)	1351	0	33220
Month	12	0	49
dew_point_2m	1019	0	32447
log_LMP	8639	0	33790
precipitation	106	0	5545
rain	122	0	5093
relative_humidity_2m	30716	0	33753
snow_depth	30	0	270
snowfall	26	0	614
temperature_2m	1092	0	33022
weather_code	13	0	11307
wind_direction_10m	8017	0	33465
wind_gusts_10m	220	0	30855
wind_speed_10m	4114	0	33500

Data Integration can be seen in given Python Code



We detected outliers using IQR Method in previous slides

Data Transformation/Feature Engineering

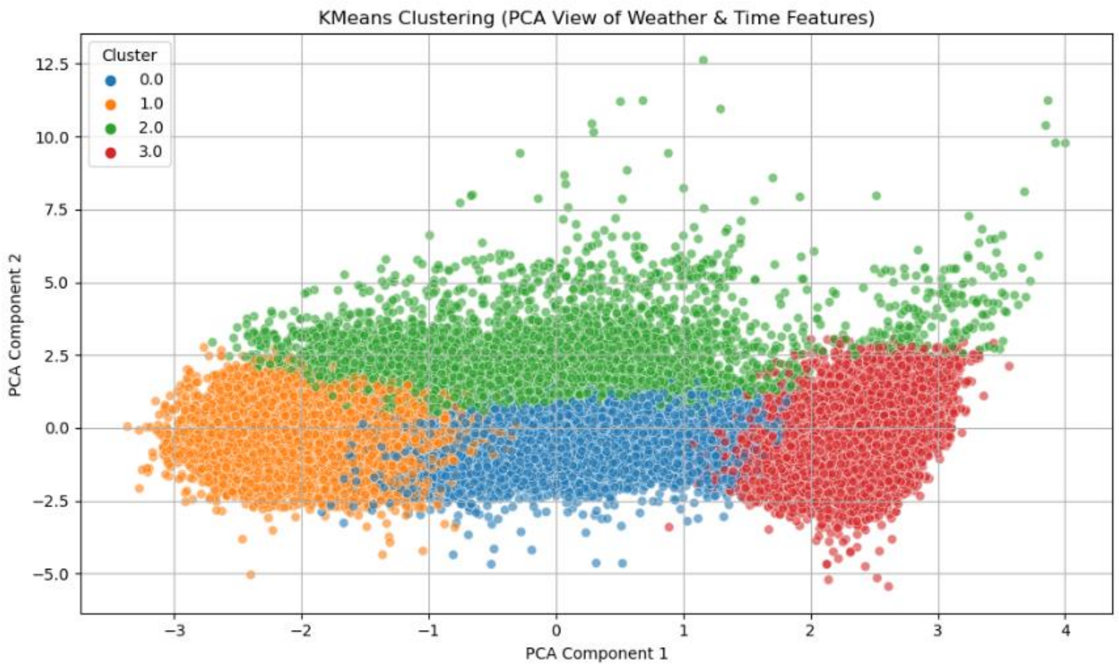
- **Attribute/feature Construction:** Used to boost our model's ability to learn from weather, time and historical trends.
  - **IsPeakHour**, Flag for peak hours (7–9 AM, 4–7 PM)
  - **IsNightHour**, Flag for nighttime hours (12–5 AM)
  - **Temp_humidity_index**, Combined weather effect (temp × humidity)
  - **Wind_total**, Wind speed + gusts
  - **IsSnowing, IsRaining**, Binary flags for precipitation types
  - **Hour_sin, Hour_cos**, Cyclical encoding of hour (captures seasonality)
  - **LMP_lag_1**, LMP value from 1 hour before
  - **LMP_lag_24**, LMP value from same hour the previous day

	IsPeakHour	IsNightHour	temp_humidity_index	wind_total \
count	33828.000000	33828.000000	33828.000000	33828.000000
mean	0.291593	0.249970	3650.750785	22.992337
std	0.454503	0.433002	1512.389387	10.841310
min	0.000000	0.000000	-601.639470	1.211160
25%	0.000000	0.000000	2510.502613	14.390856
50%	0.000000	0.000000	3546.708378	21.207304
75%	1.000000	0.000000	4789.725133	29.559548
max	1.000000	1.000000	7640.563447	84.330063

	IsSnowing	IsRaining	Hour_sin	Hour_cos	LMP_lag_1 \
count	33828.000000	33828.000000	33828.000000	3.382800e+04	33828.000000
mean	0.017500	0.133055	-0.000130	-4.751113e-05	42.060846
std	0.131128	0.339640	0.707038	7.071963e-01	41.524726
min	0.000000	0.000000	-1.000000	-1.000000e+00	0.710000
25%	0.000000	0.000000	-0.707107	-7.071068e-01	23.770000
50%	0.000000	0.000000	0.000000	-1.836970e-16	31.430000
75%	0.000000	0.000000	0.707107	7.071068e-01	48.130000
max	1.000000	1.000000	1.000000	1.000000e+00	2280.330000

	LMP_lag_24
count	33828.000000
mean	42.055544
std	41.518461
min	0.710000
25%	23.770000
50%	31.430000
75%	48.120000
max	2280.330000

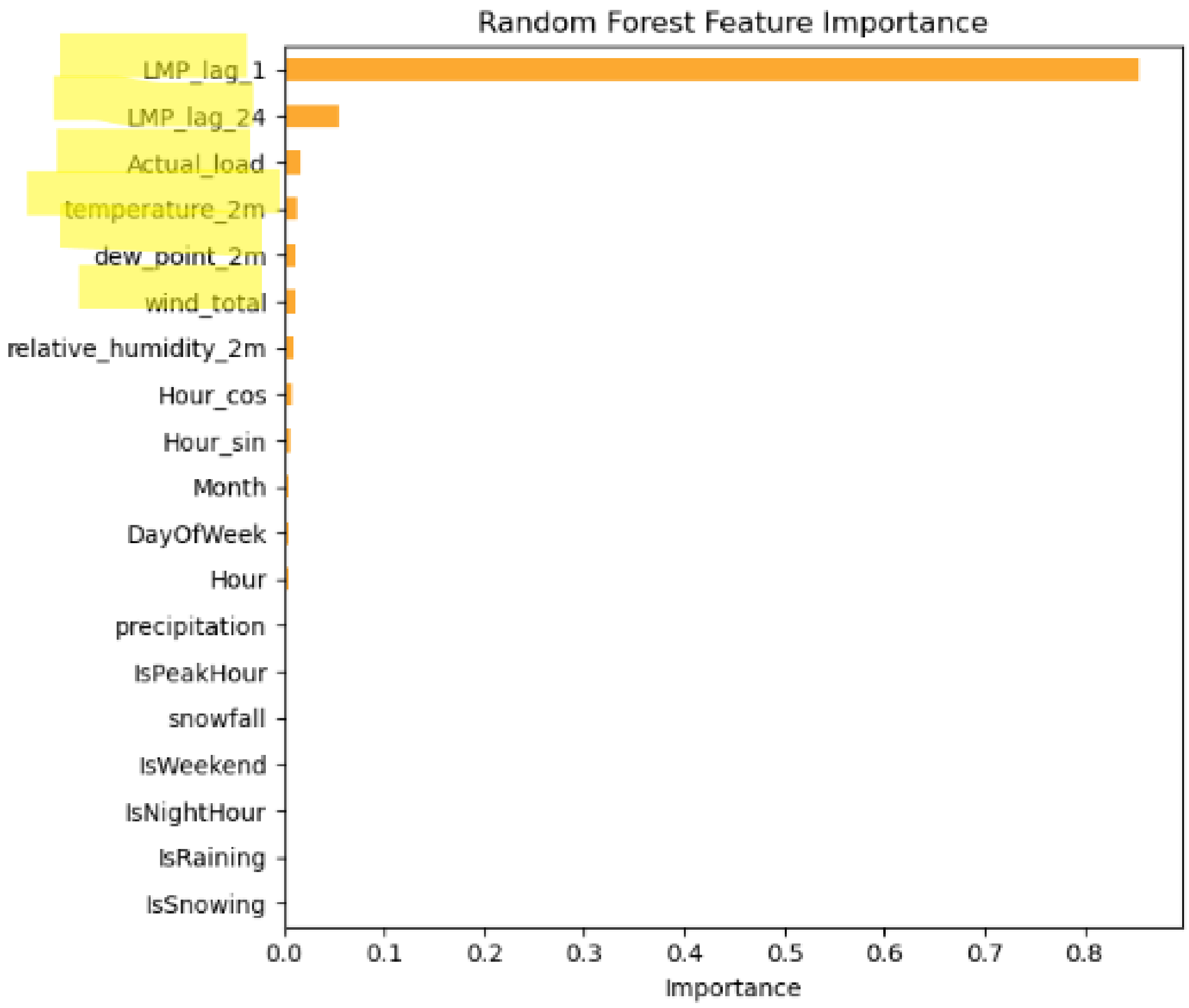
Clustering (KMeans) Unsupervised MLA

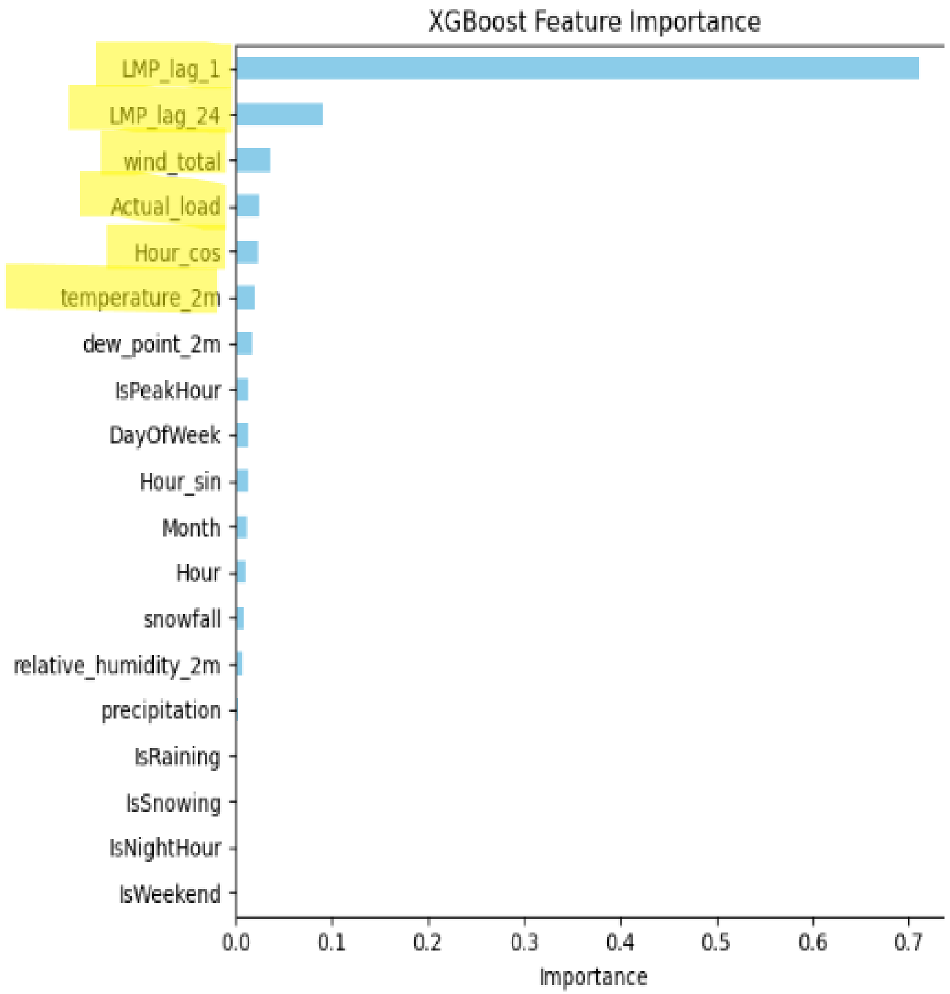


- **Group similar hours/days/weather conditions based on:**
- Weather variables
- Time of day
- LMP behavior
- Engineered Features

Cluster	Visual Shape/Area	Interpretation	LMP Behavior
0	Middle/low spread (blue)	Moderate Conditions	Medium LMP
1	(orange) Cluster	Off-peak	Low LMP
2	(green) Vertical spread	More weather variability-snow/high winds	Wide LMP Range
3	(red) Cluster	Warm hours, peak	Higher LMP

KMeans grouped ~34,000 hours into **4 clusters** based on similarity across those features.





Feature Importance for Model

Feature Reduction Example

Random Forest Regressor

Model 1 Forecast-Only (No Congestion or Loss)
RMSE: $26.61/MWh

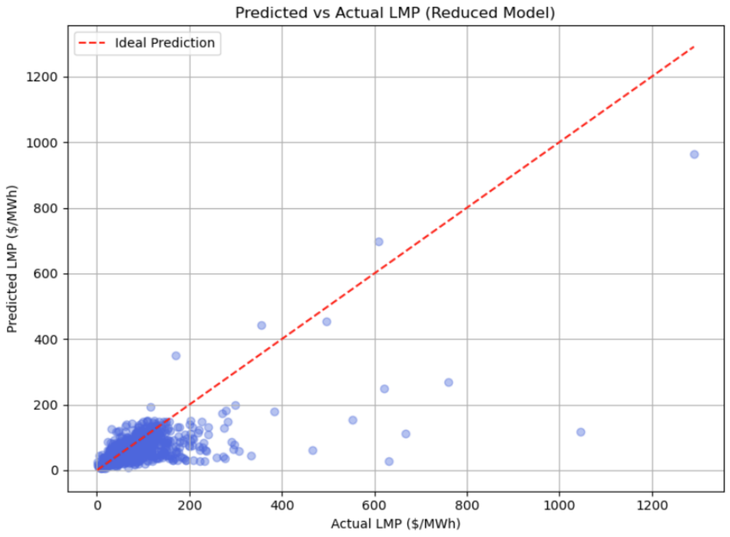
Model 1 Feature Reduction Steps and Results

Cross Validation RMSE (log LMP) 5 Folds: [26.5748709, 24.37215679, 32.29395324, 27.5589869, 36.11620113]

Select Top Features: ['temperature_2m', 'dew_point_2m', 'wind_total', 'LMP_lag_1', 'LMP_lag_24', 'Actual_load']

Retrained Reduced Model with Top Features RMSE: 26.63/MWh

Retrained Reduced Model R^2: 0.5689



5-Fold Validation

- Data split into 5 equal parts
- Model trained on 4 folds and tested on remaining 1
- This is repeated 5 times, so each fold serves as the test set once

Feature Reduction Example

Random Forest Regressor

Model 2 Full Reconstruction (with Cong. And Loss)
RMSE: $20.28/MWh
R^2 Score: 0.7500

Model 2 Feature Reduction Steps and Results

Cross Validation RMSE (log LMP) 5 Folds: [16.6854429, 21.82165671, 22.29315055, 19.3296461, 18.2396887]

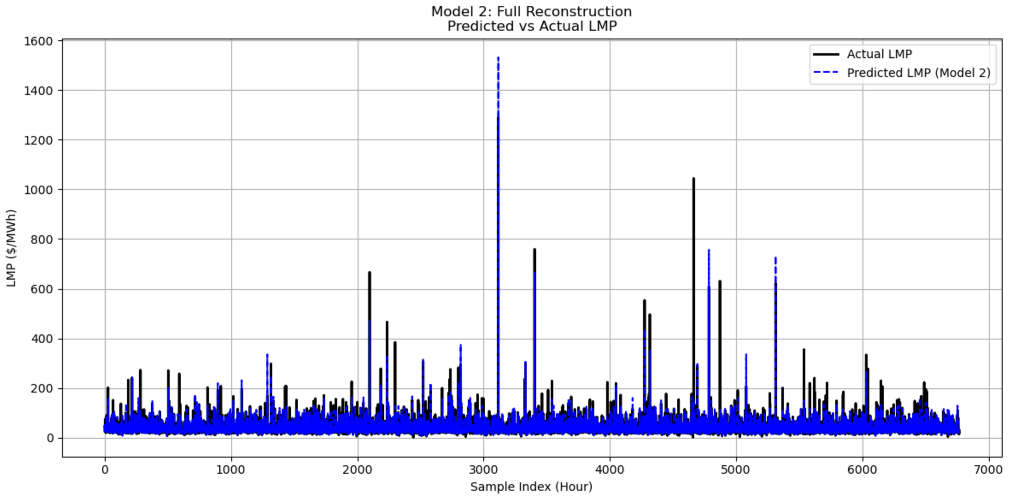
Select Top Features: [‘Michigan Hub (Loss)', ‘Michigan Hub (Congestion)', 'LMP_lag_1', 'LMP_lag_24', 'Actual_load’, ‘temperature_2m’]

Retrained Reduced Model with Top Features RMSE: 20.28/MWh

Retrained Reduced Model R^2: 0.7500

5-Fold Validation

- Data split into 5 equal parts
- Model trained on 4 folds and tested on remaining 1
- This is repeated 5 times, so each fold serves as the test set once

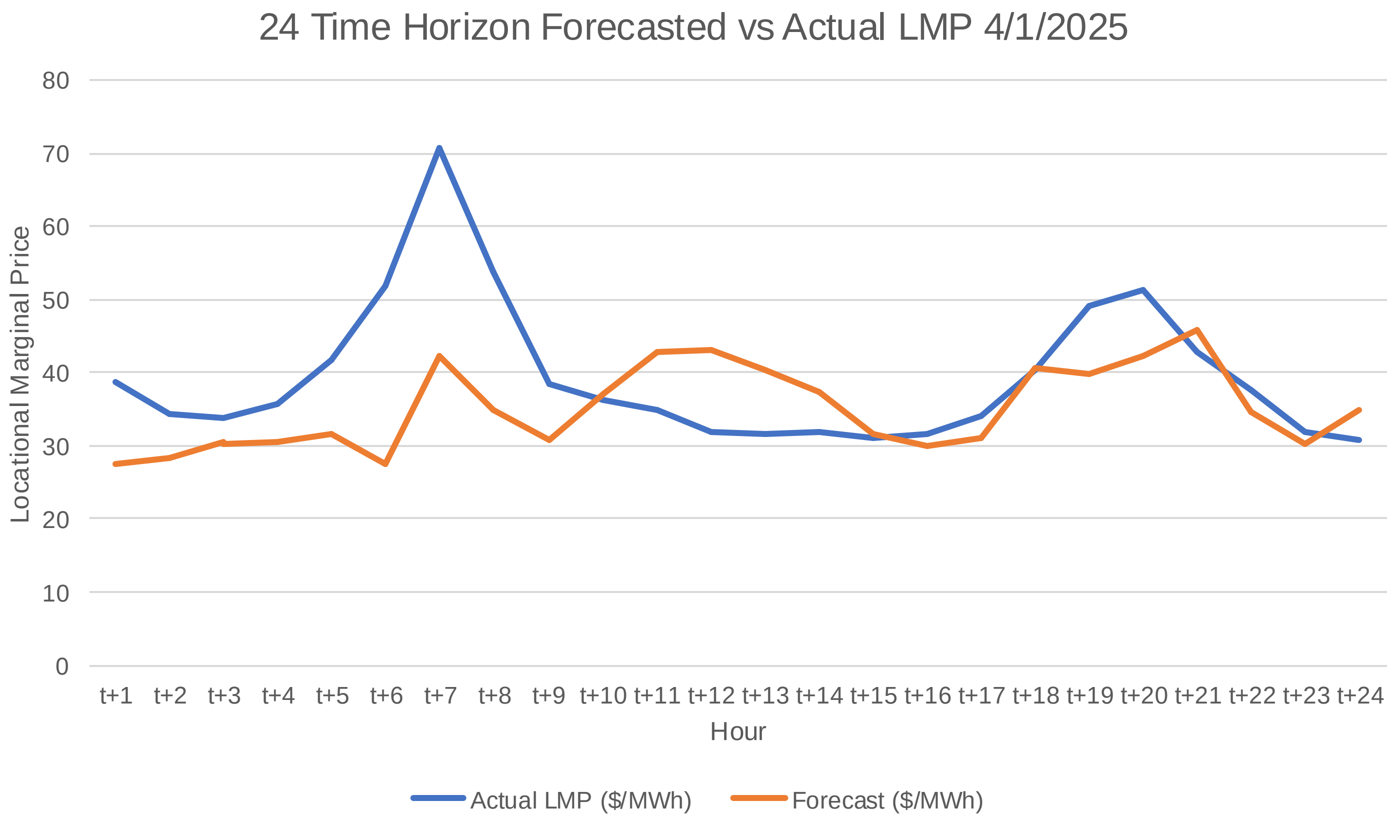


Final Model

Random Forest Regressor

Model 3 Forecast next 24 Hours 4/1/2025 (Model 2/10/21-3/31/2025)

Metric	Value
Mean Absolute Error (MAE)	$7.71/MWh
Root Mean Squared Error (RMSE)	$10.51/MWh
Mean Absolute Percentage Error(MAPE)	17.87%
R^2	.689



References

References

[1] B. Gołębiewska and J. Trajer, "Analysis of energy market using data mining methods." [Online]. Available: www.cire.pl

[2] K. R. Jay Rosano and A. C. Nerves, "Give to AgEcon Search Forecasting Locational Marginal Prices in Electricity Markets by Using Artificial Neural Networks." [Online]. Available: http://ageconsearch.umn.edu

[3] Francisco Martínez-Álvarez, Alicia Troncoso, "A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting," Energies (Basel), vol. 8, no. 11, pp. 13096–13111, 2015, doi: 10.3390/en81112361.

Thank You!!!